Prospective Analysis of Case-Control Data under General Multiplicative-Intercept Risk Models

Clarice R. Weinberg; Sholom Wacholder

Biometrika, Vol. 80, No. 2 (Jun., 1993), 461-465.

Biometrika (1993), **80**, 2, *pp.* 461-5
Printed in Great Britain

Prospective analysis of case-control data under general multiplicative-intercept risk models

By CLARICE R. WEINBERG

MD A3-03, Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, P.O. 12233, Research Triangle Park, North Carolina 27709, U.S.A.

AND SHOLOM WACHOLDER

National Cancer Institute, Executive Plaza North, Room 403, 6130 Executive Boulevard, Rockville, Maryland 20852, U.S.A.

Summary

We show that, under a general multiplicative-intercept model for risk, case-control data can be analyzed by maximum likelihood as if they had arisen prospectively, up to an unidentifiable multiplicative constant which depends on the relative sampling fractions. This generalizes earlier work of Anderson (1972, 1979), by showing that not only the point estimates but also the standard errors based on the observed information matrix are correct when the prospective likelihood is maximized. Likelihood ratio testing is also valid under this broad class of risk models. Data on disease status from a much larger cohort are shown to add no information to the estimation of the covariate-related parameters.

Some key words: Case-control study; Choice-based sampling; Cohort study; Epidemiology; Expectation-maximization, EM algorithm; Retrospective study.

1. Introduction

Anderson (1972) assumed a logistic model for risk of disease and showed that categorical data arising from retrospective studies of independently sampled diseased and nondiseased individuals can be analyzed as if the data had been collected in a prospective study, where one artificially models the disease outcome as if binomially distributed, conditional on the covariates. Prentice & Pyke (1979) later generalized this result to allow for continuous covariates, and Scott & Wild (1989) showed that likelihood ratio testing is also valid in the same 'choice-based' sampling scenario. Anderson (1979) showed that for discrete covariates, under very general assumptions about the risk model, the maximization of the prospective likelihood produces the same point estimates for odds ratios as the constrained maximization of the correct, retrospective likelihood. The purpose of this paper is to develop an extension of these results to show that the prospective analysis leads not only to valid point estimates, but to valid estimates of standard errors and likelihood ratio testing, under the broad class of 'multiplicative-intercept' models described by Hsieh, Manski & McFadden (1985).

2. Approach

We shall assume a general multiplicative-intercept model for risk, so that

$$P(D|z)/\{1-P(D|z)\}=hf(z;\beta),$$

where $P(D|z)$ is the probability of disease given covariate vector z, f is, for fixed z, a twice-differentiable function from $\mathfrak{R}^p$ to $\mathfrak{R}^+$, h is an unknown positive scalar, and β is a p-vector of unknown parameters. Note that linearity is not assumed, but z will be assumed discrete. For notational simplicity, we take disease status to be dichotomous, but generalization to polytomous models is straightforward.

Our approach involves embedding the case-control study within the implicit but unobserved larger population from which the study subjects were sampled, and treating the hypothetical resulting data structure as a missing data problem. If we had access to the whole cohort, the sampling structure would be that of a prospective study. With case-control sampling, we instead generate separate random samples for cases and for controls.

The full-data likelihood based on the unobserved cohort is developed as follows. Let n_D denote the number of cases with covariate status determined, i.e. sampled for the case-control study, and n_C denote the number of controls. Let $n_D(z)$ and $n_C(z)$ denote the respective numbers of cases and controls found to be at value z of the covariate. Suppose an additional number of cases, m_D, and noncases, m_C, have occurred in the larger cohort, but are missing the z information. We let $N = m_D + m_C + n_D + n_C$. Then we have the following hypothetical full-data likelihood appropriate to the unobserved cohort:

$$\{\textstyle\prod_z P(z|D)^{n_D(z)} P(z|\bar{D})^{n_C(z)}\} P(D)^{(n_D+m_D)} P(\bar{D})^{(n_C+m_C)}. \tag{1}$$

Using the Bayesian inversion, $P(z|D) = P(D|z)P(z)/P(D)$, (1) can be rewritten as

$$\{\textstyle\prod_z P(D|z)^{n_D(z)} P(\bar{D}|z)^{n_C(z)}\}\{\textstyle\prod_z P(z)^{(n_D(z)+n_C(z))}\}\{P(D)^{m_D} P(\bar{D})^{m_C}\} = L_1 L_2 L_3.$$

This is similar to the likelihood described by Scott & Wild (1991) in the context of stratified sampling in a case-control study. Note that factor L_1, the prospective part of the likelihood, depends only on the risk parameters, factor L_2 depends only on the nuisance parameters associated with the distribution of Z, and factor L_3 depends on both. The question is: Can proper inference be achieved by simply maximizing factor L_1, that is, using only the retrospective data modelled as if prospective? Our strategy will be to show that the two likelihoods, the improper prospective likelihood based on the case-control data and the proper full-data likelihood, lead to the same estimators for β, with the same asymptotic variance estimators.

3. MAIN RESULT

PROPOSITION 1. *Assume* $P(D|z)/\{1-P(D|z)\} = hf(z; \beta)$. *Let* $(\hat{h}^*, \hat{\beta}^*)$ *denote the solutions to the likelihood equations based on* L_1, *and* $(\hat{h}, \hat{\beta})$ *denote the solutions based on the full-data likelihood,* $L_1 L_2 L_3$, *where the distribution* $P(z)$ *is not parametrically specified. Then*

$$\hat{h} = \hat{h}^*[\{(n_D + m_D)n_C\}/\{n_D(n_C + m_C)\}].$$

The two log profile likelihoods constructed as functions of β *are parallel*; *thus* $\hat{\beta} = \hat{\beta}^*$ *and the two estimated asymptotic variance-covariance matrices for* $\hat{\beta}$ *and* $\hat{\beta}^*$ *based on the observed informations coincide.*

Proof. The proof relies on the fact that the maximum likelihood estimates for the full data occur at a stationary point of the Expectation-Maximization (EM) algorithm; see Dempster, Laird & Rubin (1977) and also Wu (1983). Let L denote the full-data likelihood, with all three of its factors, where we include data from the hypothetical cohort in which the case-control data are embedded. For the unobservable complete-data likelihood, where each subject provides complete covariate information, the sufficient statistics would be the total counts at each level of z, $N_D(z)$ and $N_C(z)$. Using instead the full data we must in the E step estimate the conditional expected frequencies at each z; the estimate for $N_D(z)$ is

$$E\{N_D(z)|n_D(z), m_D, \hat{\beta}, \hat{h}\} = n_D(z) + m_D P(z|D, \hat{\beta}, \hat{h}),$$

with a similar expression for $N_C(z)$. The M step then plugs in these estimated frequencies and maximizes the L_1 form of the likelihood (1) over h and β. The estimate for $P(z)$ is

$$\hat{P}(z)=\frac{n_D(z)+n_C(z)}{N-[\hat{P}(D\,|\,z)m_D/\hat{P}(D)+\{1-\hat{P}(D\,|\,z)\}m_C/\hat{P}(\bar{D})]} \tag{2}$$

and these can be shown to sum to 1.

Next consider the case-control data. Let $P^*(D\,|\,z)$ denote the fitted probability based on solving the likelihood equation for h that involves only L_1, with β fixed, so that

$$P^*(D\,|\,z)=\hat{h}^*f(z;\beta)/\{1+\hat{h}^*f(z;\beta)\}.$$

Maximizing over h leads to:

$$\textstyle\sum_z\{n_D(z)+n_C(z)\}P^*(D\,|\,z)=\sum_z n_D(z)=n_D. \tag{3}$$

Thus, the fitted numbers of cases must sum to the observed total number of cases. Uniqueness of $\hat{h}^*$ follows from monotonicity of $P(D\,|\,z)=hf(z;\beta)/\{1+hf(z;\beta)\}$ in h.

A similar equality holds for the full data, with the estimated counts at each z appearing in place of $n_D(z)+n_C(z)$ in (3), when estimating h so that, for any fixed β, the maximum likelihood estimate, $\hat{h}$, is selected so that the fitted $P(D)$ equals the observed proportion with disease, $(n_D+m_D)/N$.

Using this fact, together with (2) and the observation that if $r=h/h^*$ then $P(D\,|\,z)/P^*(D\,|\,z)=r-(r-1)P(D\,|\,z)$, one can show by solving for r that, for β fixed,

$$\hat{P}(z)=\frac{\{n_D(z)+n_C(z)\}\hat{P}^*(D\,|\,z)}{\{N-m_D/\hat{P}(D)\}\hat{P}(D\,|\,z)} \tag{4}$$

if and only if

$$r=\{(n_D+m_D)n_C\}/\{n_D(n_C+m_C)\}.$$

Thus $\hat{h}^*=r\hat{h}$ solves the likelihood equation involving L_1, that is equation (3), if $\hat{h}$ solves the analogous full-data likelihood equation.

We next show that the two log profile likelihoods are parallel, which will imply both that $\hat{\beta}=\hat{\beta}^*$ and that the estimated asymptotic variance matrix for $\hat{\beta}$ based on the observed information in the case-control analysis is identical to that based on the full-data likelihood.

Let $S(\beta,\hat{h}(\beta))$ denote the profile likelihood function that, for fixed full data and β, maximizes the likelihood over all choices of h and $P(z)$. Let $S^*(\beta,\hat{h}^*(\beta))$ denote the analogous profile likelihood based on the case-control data and the prospective likelihood.

Consider $\ln\{S(\beta,\hat{h}(\beta))\}-\ln\{S^*(\beta,\hat{h}^*(\beta))\}$. For each β, under the full-data likelihood $\hat{h}(\beta)$ must be set so that $\hat{P}(D)=(m_D+n_D)/N$. Thus $\hat{P}(D)$ depends only on the data and not on β. Thus in considering the difference we can restrict attention to

$$\textstyle\sum_z\left[n_D(z)\ln\left\{\dfrac{\hat{P}(D\,|\,z)}{\hat{P}^*(D\,|\,z)}\right\}+n_C(z)\ln\left\{\dfrac{\hat{P}(\bar{D}\,|\,z)}{\hat{P}^*(\bar{D}\,|\,z)}\right\}+\{n_D(z)+n_C(z)\}\ln\{P(z)\}\right]. \tag{5}$$

One can algebraically show the following two identities

$$\ln\left\{\frac{\hat{P}(D\,|\,z)}{\hat{P}^*(D\,|\,z)}\right\}=\ln\{r-(r-1)\hat{P}(D\,|\,z)\},$$

$$\ln\left\{\frac{\hat{P}(\bar{D}\,|\,z)}{\hat{P}^*(\bar{D}\,|\,z)}\right\}=-\ln(r)+\ln\{r-(r-1)\hat{P}(D\,|\,z)\},$$

and from (4)

$$\ln\{\hat{P}(z)\}=\ln\{n_D(z)+n_C(z)\}-\ln\left\{N-\frac{m_D}{\hat{P}(D)}\right\}-\ln\{r-(r-1)\hat{P}(D\,|\,z)\}$$

for r as in Proposition 1. Since r does not depend on β, substitution into (5) yields

$$\sum_z \left[-n_C(z) \ln(r) + \{n_D(z)+n_C(z)\} \ln\{n_D(z)+n_C(z)\} - \{n_D(z)+n_C(z)\} \ln\left\{N - \frac{m_D}{\hat{P}(D)}\right\}\right].$$

This depends on the data, but not on the parameter β. The log profile likelihoods are therefore parallel, implying (Richards, 1961) that the two estimated asymptotic variances for $\hat{\beta}$ based on the observed information matrices are identical. This completes the proof. □

4. Related results

Some corollaries of this result and its proof follow straightforwardly. Generalization to polytomous outcome data is quite direct, and demonstration of this will be left to the reader. We only note that the case-control estimated risk function is again related to the full-data estimates via the sampling fractions, as follows:

$$\hat{P}^*(D_i \mid z) = \frac{s_i \hat{P}(D_i \mid z)}{\sum_j s_j \hat{P}(D_j \mid z)}$$

where $s_i = n_{D_i}/(n_{D_i} + m_{D_i})$.

Although we initially posited the existence of a larger embedding cohort as a device to aid in the proof, it might happen that in certain circumstances we actually have access to the marginal disease status data from a larger, more complete population. One might ask whether the resulting improvement in precision of estimation of the background disease rate can improve precision of estimation for the β parameters. The answer is No, as stated in the following Corollary.

COROLLARY 1. *Data on disease status from a larger cohort does not add information to the estimation of β under any multiplicative-intercept model for risk.*

Thus, one cannot improve precision by bringing in information from the larger cohort, unless that information includes data on the covariates of primary interest.

Finally, we have the following Corollary, which generalizes the result of Scott & Wild (1989), at least for the case where Z is discrete.

COROLLARY 2. *Likelihood ratio testing based on the prospective analysis of case-control data is valid for multiplicative-intercept models, for tests involving parameters in β.*

Proof. The proof follows immediately from the proof of Proposition 1, since the difference between the logarithm of the maximized full-data likelihood and that of the maximized case-control data likelihood depends on the data, but not on β. Since likelihood ratio tests are valid for the full-data likelihood, large-sample validity for likelihood ratio testing based on the case-control likelihood follows. □

5. Discussion

Although identifiability is obvious for the special case of the linear logistic model, there is no guarantee that the β parameter will be identifiable in general. Consider, for example, an odds-additive model, where $P(D \mid z)/\{1 - P(D \mid z)\}$ is $z\beta$. This sort of model, where $f(z;\beta)$ is simply linear, can easily be fitted in GLIM. If the disease is rare, as is typically the situation in a case-control study, then this is approximately equivalent to simple linearity for $P(D \mid z)$. Such a model may be of particular interest in aetiological investigations since it implies approximate additivity of effect for the modelled factors. It is clear from Proposition 1 that under this model β can be specified only up to an unknown multiplicative factor, so that we can only hope to be able to estimate the relative effect ratios, β_i/β_j, for the components of β. Likelihood ratio testing would, however, still be valid for comparing nested additive models. Since independent action of two exposures is often modelled by additivity in their joint effect, aetiological hypotheses related to synergism can thus be tested using case-control data. See Wacholder & Weinberg (1993) for an example.

Acknowledgements

The authors wish to thank Drs R. Carroll, M. Gail, G. Dinse, and W. Piegorsch for helpful comments.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Anderson, J. A. (1979). Robust inference using logistic models. *Int. Statist. Inst. Bull.* **48**, 35–53.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B **39**, 1–38.

Hsieh, D. A., Manski, C. F. & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Am. Statist. Assoc.* **80**, 651–62.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *J. R. Statist. Soc.* B **23**, 469–75.

Scott, A. J. & Wild, C. J. (1989). Hypothesis testing in case-control studies. *Biometrika* **76**, 806–8.

Scott, A. J. & Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* **47**, 497–510.

Wacholder, S. & Weinberg, C. R. (1993). Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. *Biometrics.* To appear.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95–103.

[*Received September* 1991. *Revised October* 1992]